

## PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-331484

(43)Date of publication of application : 30.11.2001

(51)Int.Cl.

G06F 17/28

G06F 17/30

(21)Application number : 2000-149413

(71)Applicant : HITACHI LTD

(22)Date of filing : 22.05.2000

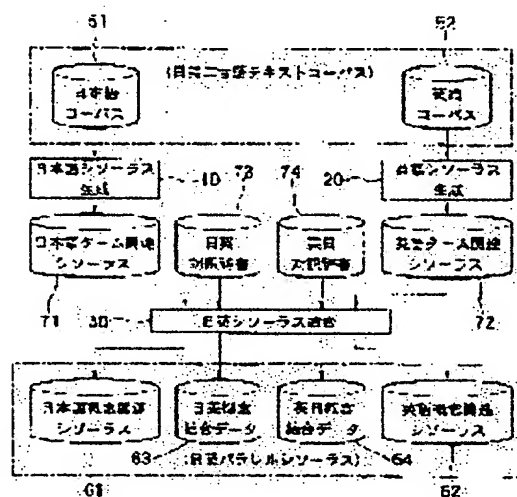
(72)Inventor : KAJI HIROYUKI  
MORIMOTO YASUTSUGU

(54) RECORDING MEDIUM HAVING PARALLEL THESAURUS GENERATION PROGRAM RECORDED THEREON, RECORDING MEDIUM HAVING PARALLEL THESAURUSES RECORDED THEREON AND RECORDING MEDIUM HAVING PARALLEL THESAURUS NAVIGATION PROGRAM RECORDED THEREON

(57)Abstract:

PROBLEM TO BE SOLVED: To automatically generate parallel thesauruses composed of concept relations within a language and concept connections between the languages and to perform text mining utilizing the generated parallel thesauruses.

SOLUTION: A server computer 1 is provided with Japanese thesaurus generation 10 for generating a Japanese term relation thesaurus 71 by extracting terms from a Japanese corpus 51 and analyzing correlations between the terms, English thesaurus generation 20 for generating English term relation thesaurus 72 by extracting the terms from an English corpus 52 and analyzing the correlations between the terms and Japanese/English thesaurus connection 30 for connecting the Japanese term relation thesaurus 71 and the English term relation thesaurus 72.



## LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2001-331484

(P2001-331484A)

(43) 公開日 平成13年11月30日 (2001. 11. 30)

(51) Int.Cl. <sup>7</sup>	識別記号	F I	タームコード* (参考)
G 0 6 F 17/28		G 0 6 F 17/28	C 5 B 0 7 5
17/30	1 7 0	17/30	1 7 0 A 5 B 0 9 1
	3 2 0		3 2 0 D
	3 5 0		3 5 0 C

審査請求 未請求 請求項の数 5 O L (全 14 頁)

(21) 出願番号 特願2000-149413 (P2000-149413)

(22) 出願日 平成12年 5 月22日 (2000. 5. 22)

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目 6 番地

(72) 発明者 梶 博行

東京都国分寺市東恋ヶ窪一丁目280番地

株式会社日立製作所中央研究所内

(72) 発明者 森本 康嗣

東京都国分寺市東恋ヶ窪一丁目280番地

株式会社日立製作所中央研究所内

(74) 代理人 100091096

弁理士 平木 祐輔

F ターム (参考) 5B075 ND03 PQ02 QM07 QP03 UU01

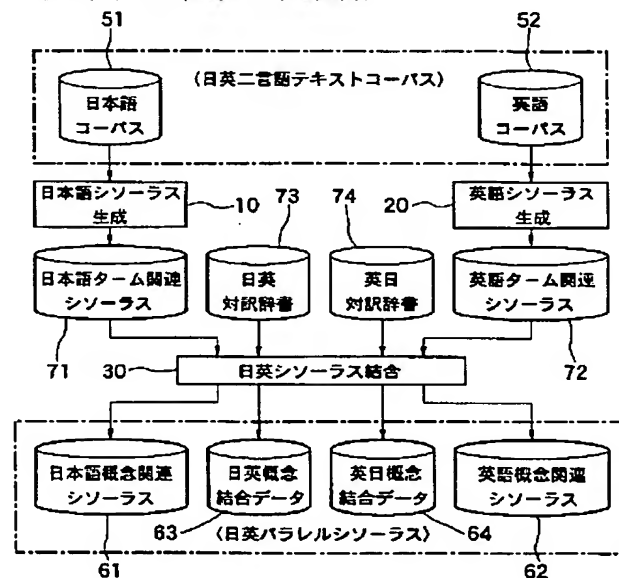
5B091 AA05 AB17 CA12 CB22

(54) 【発明の名称】 パラレルシソーラスの生成プログラムを記録した記録媒体、パラレルシソーラスを記録した記録媒体及びパラレルシソーラスナビゲーションプログラムを記録した記録媒体

(57) 【要約】

【課題】 言語内の概念関連と言語間の概念結合からなるパラレルシソーラスを自動的に生成する。生成したパラレルシソーラスを利用したテキストマイニングを実現する。

【解決手段】 サーバ計算機 1 は、日本語コーパス 5 1 からタームを抽出し、ターム間の相関を解析することにより日本語ターム関連シソーラス 7 1 を生成する日本語シソーラス生成 1 0 と、英語コーパス 5 2 からタームを抽出し、ターム間の相関を解析することにより英語ターム関連シソーラス 7 2 を生成する英語シソーラス生成 2 0 と、日本語ターム関連シソーラス 7 1 と英語ターム関連シソーラス 7 2 とを結合する日英シソーラス結合 3 0 とを備える。



**【特許請求の範囲】**

**【請求項 1】** コンピュータを、第 1 言語のテキストコーパスからタームを抽出し、ターム間の相関を解析することにより第 1 言語のターム関連シソーラスを生成する第 1 言語シソーラス生成手段と、第 2 言語のテキストコーパスからタームを抽出し、ターム間の相関を解析することにより第 2 言語のターム関連シソーラスを生成する第 2 言語シソーラス生成手段と、第 1 言語のターム関連シソーラスと第 2 言語のターム関連シソーラスを結合するシソーラス結合手段と、を備えるパラレルシソーラスの生成装置として機能させるためのプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

**【請求項 2】** 前記シソーラス結合手段は、対訳辞書を参照して第 1 言語のターム関連シソーラスと第 2 言語のターム関連シソーラスとの間で対応するタームを結合するターム結合手段と、結合された第 2 言語のタームを組み合わせることにより第 1 言語の各タームから概念ラベルを生成する第 1 言語概念ラベル生成手段と、結合された第 1 言語のタームを組み合わせることにより第 2 言語の各タームから概念ラベルを生成する第 2 言語概念ラベル生成手段と、言語間のターム結合を言語間の概念結合に変換する概念結合手段と、第 1 言語のターム関連シソーラスに含まれるターム間の関連を概念間の関連に変換する第 1 言語概念関連シソーラス生成手段と、第 2 言語のターム関連シソーラスに含まれるターム間の関連を概念間の関連に変換する第 2 言語概念関連シソーラス生成手段と、同一の第 2 言語の概念に結合され、関連する第 1 言語の概念の集合が類似している第 1 言語の概念をマージして一つの概念にする第 1 言語概念マージ手段と、同一の第 1 言語の概念に結合され、関連する第 2 言語の概念の集合が類似している第 2 言語の概念をマージして一つの概念にする第 2 言語概念マージ手段と、を有することを特徴とする請求項 1 記載のコンピュータ読み取り可能な記録媒体。

**【請求項 3】** 第 1 言語のタームと第 2 言語のタームを組み合わせた概念ラベルと概念ラベルが表す概念に基づく概念ラベルの結合とから構成されるシソーラスを記録したことを特徴とするパラレルシソーラスを記録した記録媒体。

**【請求項 4】** コンピュータを、第 1 言語の複数のターム又は概念の集合から、第 2 言語の複数のターム又は概念の集合に遷移する遷移手段を備えるパラレルシソーラスナビゲーションシステムとして機能させるためのプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

**【請求項 5】** 前記遷移手段は、第 1 言語のターム又は概念の集合から、該集合中のターム又は概念に結合された第 2 言語のターム又は概念に、該第 2 言語のターム又は概念の関連ターム又は関連概念であって、第 1 言語の

ターム又は概念に結合されていないターム又は概念を加えた集合に遷移することを特徴とする請求項 4 記載のコンピュータ読み取り可能な記録媒体。

**【発明の詳細な説明】****【0001】**

**【発明の属する技術分野】** 本発明は、二つの言語のテキストコーパスから二つの言語のシソーラスを結合したパラレルシソーラスを生成する装置及び該パラレルシソーラスを利用したナビゲーションシステムのプログラム、並びにパラレルシソーラスを記録した記録媒体に関する。

**【0002】**

**【従来の技術】** 電子化されたテキスト情報の増加と共に、情報アクセス技術の重要性が高まっている。本発明者らは先に特願平 11-28101 号として、シソーラスのナビゲーション方法を出願している。この方法は、テキストデータを記憶する文書データベース（以下、テキストコーパスと呼ぶ）から、有益な情報を掘り出す作業（以下、テキストマイニングと呼ぶ）を効率的に行う技術であり、テキストコーパスからターム及びターム間の関連知識を抽出してシソーラスを生成し、該シソーラスの内容をクライアント端末のブラウザに表示することによりナビゲーションする。

**【0003】** また、情報処理学会データベースシステム研究会／情報学基礎研究会研究報告 DBS-118-13/FI-54-13「コーパス対応の関連シソーラスナビゲーション」（1999 年 5 月 17 日）には、上記ナビゲーション方法に基づくテキストマイニングシステムが報告されている。

**【0004】**

**【発明が解決しようとする課題】** 情報検索の分野では、母国語で表現した検索要求を入力して、外国語の文書を検索したいというニーズが高まっている。この情報検索方法は、クロスランゲージ情報検索と呼ばれ、盛んに研究されている。

**【0005】** クロスランゲージ情報検索の代表的な手法は、たとえば情報処理学会論文誌 40 巻 11 号ページ 4075-4086「機械翻訳を用いた英日・日英言語横断検索に関する一考察」（1999 年 11 月）に報告されている。この検索方法は、対訳辞書や機械翻訳システムを利用して、検索要求を文書と同じ言語に翻訳した上で文書検索を実行する。この場合、検索要求は文書に比べて短く文脈情報が少ないため、高精度で翻訳するのが難しく、検索精度が低いという問題がある。

**【0006】** この問題に対して、上記特願平 11-28101 号で提案しているナビゲーション方法を利用することが考えられる。この場合、単言語から二言語に拡張して情報検索システムのフロントエンドとして使用することにより、クロスランゲージ情報検索における上記問題点を解決することができると考えられるが、二つの言

語のシソーラスを結合することが必要になる。

【0007】一方、テキストコーパスからのシソーラス生成技術は、上述したようなテキストマイニングへの応用だけでなく、様々な自然言語処理応用システムに有効であるが、次のような技術課題が残されている。

【0008】従来のシソーラス自動生成の技術は、多義語の取扱いに関して問題がある。従来の技術ではターム間の関連を抽出しているが、ターム間の関連は、本来意味的なものであるもので、多義語のタームを語義すなわち概念に分割し、概念間の関連を抽出するのが理想的である。従来の技術によれば、たとえば「bank」の関連タームとして「loan」、「money」、「river」及び「water」等が、「bank」の概念を問わず全て抽出されてしまう。

【0009】ここで「loan」及び「money」は、お金を預けたり引き出したりする機関としての「bank（銀行）」の関連タームであり、「river」及び「water」は、水辺の場所としての「bank（岸）」の関連タームである。したがって、「bank」をそれが表す概念に分割し、「loan」及び「money」等の関連タームと、「river」及び「water」等の関連タームとが別々に抽出されることが望ましい。

【0010】また、従来のシソーラス自動生成技術は、同義語の取扱いに関しても問題がある。従来の技術では共起確率、すなわちテキスト中の近傍に揃って出現する確率に基づいて関連タームを抽出している。このため、同義語は関連タームとしてさえ抽出されず、別々のエンティティとして扱われる。同義語は同じ概念を表すのであるから、同義語が一つのエンティティに纏めて取扱われることが望ましい。

【0011】以上より、本発明の目的は上述した従来の技術における問題点を解決することである。第1の目的は、第1言語の概念及び概念間の関連、第2言語の概念及び概念間の関連、第1言語の概念と第2言語の概念間の結合から構成されるパラレルシソーラスを第1言語及び第2言語のテキストコーパスから自動生成する、パラレルシソーラスの生成プログラムを記録した記録媒体を提供することにある。

【0012】第2の目的は、クロスランゲージ情報検索のフロントエンドとして、特に、多義語及び同義語の取扱いに注目して、検索要求の高精度な翻訳を可能にする、パラレルシソーラスナビゲーションプログラムを記録した記録媒体を提供することにある。第3の目的は、上記パラレルシソーラスを利用した有効なテキストマイニングを実現するために、パラレルシソーラスを記録した記録媒体を提供することにある。

【0013】

【課題を解決するための手段】本発明は、コンピュータを、第1言語のテキストコーパスからタームを抽出し、ターム間の相関を解析することにより第1言語のターム

関連シソーラスを生成する第1言語シソーラス生成手段と、第2言語のテキストコーパスからタームを抽出し、ターム間の相関を解析することにより第2言語のターム関連シソーラスを生成する第2言語シソーラス生成手段と、第1言語のターム関連シソーラスと第2言語のターム関連シソーラスを結合するシソーラス結合手段と、を備えるパラレルシソーラスの生成装置として機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【0014】また、前記シソーラス結合手段は、対訳辞書を参照して第1言語のターム関連シソーラスと第2言語のターム関連シソーラスとの間で対応するタームを結合するターム結合手段と、結合された第2言語のタームを組み合わせることにより第1言語の各タームから概念ラベルを生成する第1言語概念ラベル生成手段と、結合された第1言語のタームを組み合わせることにより第2言語の各タームから概念ラベルを生成する第2言語概念ラベル生成手段と、言語間のターム結合を言語間の概念結合に変換する概念結合手段と、第1言語のターム関連シソーラスに含まれるターム間の関連を概念間の関連に変換する第1言語概念関連シソーラス生成手段と、第2言語のターム関連シソーラスに含まれるターム間の関連を概念間の関連に変換する第2言語概念関連シソーラス生成手段と、同一の第2言語の概念に結合され、関連する第1言語の概念の集合が類似している第1言語の概念をマージして一つの概念にする第1言語概念マージ手段と、同一の第1言語の概念に結合され、関連する第2言語の概念の集合が類似している第2言語の概念をマージして一つの概念にする第2言語概念マージ手段と、を有する。これにより、多義語及び同義語が有する概念を考慮したパラレルシソーラスを生成することができる。

【0015】パラレルシソーラスの生成装置は、以下のように作用する。第1言語シソーラス生成手段は、第1言語のテキストコーパスからターム及び共起するタームの組を抽出し、ターム間の相関を解析することにより、各ターム毎に関連タームの集合を出力する。第1言語が日本語であるとき、たとえば、ターム「銀行」の関連タームの集合として{ローン, 金利, 口座, 利率, 証券, 経済, 金融, 投資}が出力される。

【0016】第2言語シソーラス生成手段は、第1言語シソーラス生成手段と同様な処理を第2言語のテキストコーパスに対して実行し、各ターム毎に関連タームの集合を出力する。第2言語が英語であるとき、たとえば、ターム「bank」の関連タームの集合として、{account, river, interest, loan, boat, investment, fishing, park, economy, lake}が出力される。

【0017】シソーラス結合手段において、各手段は以下のように作用する。ターム結合手段は、対訳辞書を参照して、第1言語のターム関連シソーラスと第2言語のターム関連シソーラスとの間で対応するタームを結合す

る。たとえば、「銀行」と「bank」が結合され、「岸」と「bank」が結合される。

【0018】第1言語概念ラベル生成手段は、結合された第2言語のタームを組み合わせることにより、第1言語の各タームから少なくとも1つの概念ラベルを生成する。同様に、第2言語概念ラベル生成手段は、結合された第1言語のタームを組み合わせることにより、第2言語の各タームから少なくとも1つの概念ラベルを生成する。たとえば「bank」は「銀行」と「岸」とに結合されている。このとき、「銀行」の関連ターム集合と「岸」の関連ターム集合が似ていなければ、「銀行」と「岸」とが概念的に異なると判断され、「bank」から二つの概念ラベル「bank・銀行」、「bank・岸」が生成される。

【0019】概念結合手段は、言語間のターム結合を言語間の概念結合に変換する。たとえば、ターム結合「銀行-bank」は概念結合「銀行-bank・銀行」に変換され、ターム結合「岸-bank」は概念結合「岸-bank・岸」に変換される。

【0020】第1言語概念関連シソーラス生成手段は、第1言語のターム関連シソーラスに含まれるターム間の関連を概念間の関連に変換する。同様に、第2言語概念関連シソーラス生成手段は、第2言語のターム関連シソーラスに含まれるターム間の関連を概念間の関連に変換する。たとえば、ターム間の関連「bank-interest」は概念間の関連「bank・銀行-interest・金利/利率」に変換され、ターム間の関連「bank-river」は概念間の関連「bank・岸-river」に変換される。

【0021】第1言語概念マージ手段は、同一の第2言語の概念に結合され、関連する第1言語の概念の集合が類似している第1言語の概念を一つの概念にマージする。同様に、第2言語概念マージ手段は、同一の第1言語の概念に結合され、関連する第2言語の概念の集合が類似している第2言語の概念を一つの概念にマージする。たとえば、日本語の二つの概念「金利」と「利率」が共に英語の概念「interest・金利/利率」に結合されている。このとき、「金利」の関連概念の集合と「利率」の関連概念の集合が類似していれば、「金利」と「利率」とが一つの概念「金利-利率」にマージされる。

【0022】以上のように各機能が作用することにより、第1言語の概念及び概念間の関連から構成される第1言語の概念関連シソーラスと、第2言語の概念及び概念間の関連から構成される第2言語の概念関連シソーラスとが結合されたパラレルシソーラスが生成される。

【0023】また、本発明は、第1言語のタームと第2言語のタームを組み合わせた概念ラベルと概念ラベルが表す概念に基づく概念ラベルの結合とから構成されるパラレルシソーラスを記録した記録媒体である。

【0024】また、本発明は、コンピュータを、第1言語の複数のターム又は概念の集合から、第2言語の複数

のターム又は概念の集合に遷移する遷移手段を備えるパラレルシソーラスナビゲーションシステムとして機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【0025】また、前記遷移手段は、第1言語のターム又は概念の集合から、該集合中のターム又は概念に結合された第2言語のターム又は概念に、該第2言語のターム又は概念の関連ターム又は関連概念であって、第1言語のターム又は概念に結合されていないターム又は概念を加えた集合に遷移する。パラレルシソーラスのナビゲーションシステムにおいて、第1言語の概念の集合から第2言語の概念の集合への遷移手段は次のように作用する。

【0026】遷移前の概念集合中の概念に結合された第2言語の概念と、それらの第2言語の概念と関連が強く、かつ第1言語の概念と結合されていない第2言語の概念を併せて遷移後の概念集合を構成する。これにより、パラレルシソーラスにおいて陽に結合されていない概念を含めて、第1言語の概念集合から関連する第2言語の概念集合へ遷移することができる。

【0027】

【発明の実施の形態】以下、本発明の実施の形態を添付図面と対応して詳細に説明する。図1は、本発明の実施の形態によるパラレルシソーラスの生成装置と、該装置を含むパラレルシソーラスナビゲーションシステムの構成を説明するブロック図である。なお、本実施の形態では、二つの言語のシソーラスが結合されたパラレルシソーラスとして、日本語と英語とによるパラレルシソーラスの生成について説明する。

【0028】本実施の形態によるパラレルシソーラスナビゲーションシステム（以下、システムと呼ぶ）は、通信ネットワーク3を介して互いに接続されるサーバ計算機1とクライアント計算機2により構成される。

【0029】サーバ計算機1は、日本語のシソーラスと英語のシソーラスとを対応付けるパラレルシソーラス生成処理と、本システムの処理のうち、シソーラスの検索処理等を行う。このサーバ計算機1は、処理装置11、入力装置12及び記憶装置13により主に構成される。

【0030】処理装置11は、サーバ計算機1の全体の処理を実行する。特に、後述する図2から図4に示すパラレルシソーラスの生成に関わる各データ処理を実行する。入力装置12は、テキストコーパスの入手媒体に応じてCD-ROMドライブ、フロッピー（登録商標）ディスクドライブ等であり、テキストコーパスの入力に用いられる。

【0031】記憶装置13は、RAM、ROM、光磁気ディスクライブラリ装置（図示せず）等の記憶手段を総称しており、たとえば、サーバ計算機1の処理プログラム等はROMに固定的に格納され、サーバ計算機1の処理の過程で作成されたデータ、作業ファイル等はRAM

に一時的に格納され、さらに各言語のコーパス、シソーラス及び対訳辞書（図2参照）等は光磁気ライブラリ装置等の大容量記憶装置に格納される。

【0032】クライアント計算機2は、本システムの処理のうち、サーバ計算機1の検索結果として送信されるシソーラスの表示、ユーザとの対話処理等を行う。つぎに、図2から図4を用いてパラレルシソーラスの生成処理の詳細を説明する。

【0033】図2は、パラレルシソーラス生成装置における入出力データとモジュール構成を機能的に説明する図である。パラレルシソーラス生成装置の入力は、日本語コーパス51と英語コーパス52とが対になった日英二言語テキストコーパスである。日本語コーパス51と英語コーパス52とは同じ分野のテキストであるという条件が課せられるが、対訳である必要はない。

【0034】パラレルシソーラス生成装置の出力は、日本語概念関連シソーラス61、英語概念関連シソーラス62、日英概念結合データ63、英日概念結合データ64からなる日英パラレルシソーラスである。日英概念結合データ63と英日概念結合データ64は、情報の内容は同じでレコード形式が違うだけである。冗長ではあるが、日英シソーラス結合処理の効率を考慮して、両方を出力する。

【0035】パラレルシソーラス生成装置を構成するモジュールは、日本語シソーラス生成10、英語シソーラス生成20、及び日英シソーラス結合30である。日本語シソーラス生成10は、日本語コーパス51から日本語ターム関連シソーラス71を生成する。英語シソーラス生成20は、英語コーパス52から英語ターム関連シソーラス72を生成する。日英シソーラス結合30は、日英対訳辞書73と英日対訳辞書74を参照して、日本語ターム関連シソーラス71と英語ターム関連シソーラス72とから日本語概念関連シソーラス61、英語概念関連シソーラス62、日英概念結合データ63、及び英日概念結合データ64を生成する。日英対訳辞書73と英日対訳辞書74は、情報の内容は同じでレコード形式が違うだけである。冗長ではあるが、日英シソーラス結合処理の効率を考慮して、両方を使用する。

【0036】図3は、日本語シソーラス生成10の処理の詳細を説明する図である。図3に示すように、日本語シソーラスを生成する処理は、ターム抽出101、共起データ抽出102、及び相関解析103の3つのステップからなる。

【0037】（1）ターム抽出101

日本語コーパス51からタームを抽出して、出現頻度をカウントする。タームとしては、出現頻度が予め定めた閾値以上の名詞と複合名詞を抽出する。複合名詞は、品詞列パターンを用いたパターンマッチングによって抽出する。高頻度語の中には、特に分野に関係のない一般的な語も多い。それらは、ストップワードリストを用いて

取り除く。

【0038】このストップワードリストに関して、たとえば、“上記”を先頭要素のストップワードとすることにより、“上記システム”というような名詞句を除外できる。同様に、“全体”を末尾要素のストップワードとすることにより、“システム全体”というような名詞句を除外できる。

【0039】（2）共起データ抽出102

共起するタームの対を抽出して、共起頻度をカウントする。共起の定義としてはウィンドウ共起を採用する。すなわち、一定の幅をもったウィンドウをテキストに沿って移動させながら、各位置でのウィンドウに含まれるタームの対を抽出する。ウィンドウの幅は、たとえば機能語を除いて25タームとする。

（3）相関解析103

全てのターム対に対して統計的な相関値を計算し、予め定めた閾値以上の相関値をもつターム対を抽出する。タームの相関値としては相互情報量を用いる。

【0040】以上述べた（1）～（3）のステップの結果として、日本語ターム関連シソーラス71が得られる。日本語ターム関連シソーラス71は、日本語の各タームに対する関連ターム集合を表すレコードの集まりである。すなわち、

$$RT_J(x_i) = \{x(i, 1), x(i, 2), \dots, x(i, m_i)\} \quad (i = 1, 2, \dots, M).$$

ここで、 $x_i$ は日本語ターム、 $x(i, m_i)$ は関連ターム、 $i$ は各タームに付される番号、 $m_i$ は第 $i$ 日本語タームの関連ターム数、 $M$ は日本語タームの総数である。レコードの例を以下に示す。

【0041】 $RT_J(\text{銀行}) = \{\text{ローン, 金利, 口座, 利率, 証券, 経済, 金融, 投資}\}.$

$RT_J(\text{金利}) = \{\text{ローン, 貸出し, 預貯金, 引き上げ, 銀行}\}.$

以上、日本語シソーラス生成10の処理を説明したが、英語シソーラス生成20の処理も、日本語シソーラス生成10と全く同様である。英語シソーラスの生成処理の結果として、英語ターム関連シソーラス72が得られる。英語ターム関連シソーラス72は、英語の各タームに対する関連ターム集合を表すレコードの集まりである。すなわち、

$$RT_E(y_i) = \{y(i, 1), y(i, 2), \dots, y(i, n_i)\} \quad (i = 1, 2, \dots, N).$$

ここで、 $y_i$ は英語ターム、 $y(i, n_i)$ は関連ターム、 $i$ は各タームに付される番号、 $n_i$ は第 $i$ 英語タームの関連ターム数、 $N$ は英語タームの総数である。レコードの例を以下に示す。

$RT_E(\text{bank}) = \{\text{account, river, interest, loan, boat, investment, fishing, park, economy, lake}\}.$

$RT_E(\text{interest}) = \{\text{loan, deposit, bank, science, economy, exchange, politics}\}.$

【0042】図4は、日英シソーラス結合モジュール30の処理の詳細を説明する図である。図4に示すように、日英シソーラス結合処理は、日英ターム結合301、日本語概念ラベル生成302、英語概念ラベル生成303、日英概念結合304、日本語概念関連シソーラス生成305、英語概念関連シソーラス生成306、日本語概念マージ307、及び英語概念マージ308の8つのステップからなる。以下、これらの処理の詳細を説明する。

【0043】(1) 日英ターム結合301

日英対訳辞書73と英日対訳辞書74を参照して、日本語ターム関連シソーラス71と英語ターム関連シソーラス72の間で対応するタームを結合し、日英ターム結合データ91と英日ターム結合データ92とを出力する。日英ターム結合の入力のうち、日本語ターム関連シソーラス71と英語ターム関連シソーラス72とは既に説明したので、日英対訳辞書73と英日対訳辞書74とについて説明する。

【0044】日英対訳辞書73は、日本語の各タームに対する対訳英語ターム集合を表すレコードの集まりである。すなわち、

$$D_{JE}(a_i) = \{b(i, 1), b(i, 2), \dots, b(i, l_i)\} \quad (i = 1, 2, \dots, K).$$

ここで、 $a$  は日本語ターム、 $b$  は英語タームである。レコードの例を以下に示す。

$$D_{JE}(\text{銀行}) = \{\text{bank}\}.$$

$$D_{JE}(\text{岸}) = \{\text{bank}\}.$$

【0045】英日対訳辞書74は、英語の各タームに対する対訳日本語ターム集合を表すレコードの集まりである。すなわち、

$$D_{EJ}(b_i) = \{a(i, 1), a(i, 2), \dots, a(i, k_i)\} \quad (i = 1, 2, \dots, L).$$

ここで、 $b$  は英語ターム、 $a$  は日本語タームである。レコードの例を以下に示す。

$$D_{EJ}(\text{bank}) = \{\text{銀行}, \text{バンク}, \text{岸}\}.$$

$$D_{EJ}(\text{interest}) = \{\text{興味}, \text{金利}, \text{利率}\}.$$

【0046】次に、日英ターム結合の出力について説明する。日英ターム結合データ91と英日ターム結合データ92とは、同じ情報を異なる形式で表現したものである。冗長ではあるが、後続の処理の効率を考慮して両方を出力する。日英ターム結合データ91は、日本語タームの各々について、それに結合された英語タームの集合を表すレコードの集まりである。すなわち、

$$TL_{JE}(x_i) = \{y'(i, 1), y'(i, 2), \dots, y'(i, n'_i)\} \\ (i = 1, 2, \dots, M).$$

ここで、 $x$  は日本語ターム、 $y'$  は英語タームである。

【0047】英日ターム結合データ92は、英語タームの各々について、それに結合された日本語タームの集合を表すレコードの集まりである。すなわち、

$$TL_{EJ}(y_i) = \{x'(i, 1), x'(i, 2), \dots, x'(i, m'_i)\} \\ (i = 1, 2, \dots, N).$$

ここで、 $y$  は英語ターム、 $x'$  は日本語タームである。

【0048】日英ターム結合301のアルゴリズムは次のとおりである。

1) 日英ターム結合データ91を初期化する。すなわち、

$$TL_{JE}(x_i) \leftarrow \phi \quad (i = 1, 2, \dots, M).$$

2) 英日ターム結合データ92を初期化する。すなわち、

$$TL_{EJ}(y_i) \leftarrow \phi \quad (i = 1, 2, \dots, N).$$

【0049】3) 次の2つの条件を満足する、日本語ターム $x$ と英語ターム $y$ とを結合する。

(a) 対訳関係 $\langle x, y \rangle$ が対訳辞書によってサポートされている。

(b) 対訳関係 $\langle x, y \rangle$ のドメイン関連度 $DR(x, y)$ が予め定めた閾値以上である。

すなわち、(a)及び(b)を満足する全ての日本語ターム $x$ と英語ターム $y$ の対に関して、

$$TL_{JE}(x) \leftarrow TL_{JE}(x) \cup \{y\} \text{ 及び } TL_{EJ}(y) \leftarrow TL_{EJ}(y) \cup \{x\}.$$

を実行する。

(a)  $x$  が $k$ 個のターム $x_1, x_2, \dots, x_k$ の並び、 $y$  が $k$ 個のターム $y_1, y_2, \dots, y_k$ の並びであって、 $\{y'_1, y'_2, \dots, y'_k\} = \{y_1, y_2, \dots, y_k\}$ であるような $y'_1 (\in D_{JE}(x_1)), y'_2 (\in D_{JE}(x_2)), \dots, y'_k (\in D_{JE}(x_k))$ が存在する。

(b)  $DR(x, y) \geq \theta$ .

【0050】条件(a)により、日本語ターム $x$ と英語ターム $y$ とが対訳関係にあるかを知るために、構成要素の間での対訳関係が成立しているかが、集合におけるタームの順番を問わずにチェックされる。特に、 $k=1$ のとき $y \in D_{JE}(x)$ となる。すなわち、対訳辞書に登録されているターム対であることを意味する。 $k \geq 2$ のとき、構成要素間の対訳関係が、対訳辞書に登録されているような複合語タームの対であることを意味する。条件(b)により、対訳辞書が示唆する対訳関係がドメインで成立する関係であるかがチェックされる。対訳関係 $\langle x, y \rangle$ のドメイン関連度 $DR(x, y)$ は次式で定義される。

【0051】

【数1】

$$DR(x, y) = \max \{ DR_{JE}(x, y), DR_{EJ}(x, y) \}.$$

$$DR_{JE}(x, y) = | \bigcup_{x' \in RT_J(x)} D_{JE}(x') \cap RT_E(y) | / | RT_J(x) |.$$

$$DR_{EJ}(x, y) = | \bigcup_{y' \in RT_E(y)} D_{EJ}(y') \cap RT_J(x) | / | RT_E(y) |.$$

【0052】 $DR_{JE}(x, y)$ は、日本語タームの関連タームのうち、英語訳が英語タームの関連タームであるものの比率である。すなわち、ある日本語タームがどのような文脈で出現するかを示す出現文脈が、英語タームの出現文脈と重なる度合である。また、 $DR_{EJ}(x, y)$ は、英語タームの出現文脈が日本語タームの出現文脈と重なる度合である。多少なりとも出現文脈に共通性があれば、対訳関係がドメインで成立すると考えてよいので、ドメイン関連度の閾値 $\theta$ は小さめに設定するのがよい。以上のアルゴリズムにより、日英ターム結合301が実行される。

【0053】(2) 日本語概念ラベル生成302

(3) 英語概念ラベル生成303

日本語概念ラベル生成302と英語概念ラベル生成303の処理は、日本語と英語の役割が反転する以外は全く同様である。したがって、ここでは英語概念ラベル生成303について説明する。

【0054】英語概念ラベル生成ステップ303は、英日ターム結合データ92と日本語ターム関連シソーラス71とに基づいて、英語タームの各々から一つ以上の英語概念ラベルを生成する。さらに、生成した英語概念ラベルの各々に対して関連タームの集合を生成する。このために、英語ターム関連シソーラス72、日英ターム結合データ91、日本語ターム関連シソーラス71、英日ターム結合データ92を参照する。

【0055】英語概念ラベル生成303の入力データは既に説明済みであるので、出力データを説明する。英語概念ラベルは、タームを組み合わせたものであり、以下のように定義される。

<英語概念ラベル> : = <英語ターム> | <英語ターム> · <日本語修飾子> | <英語概念ラベル> + <英語概念ラベル>.

<日本語修飾子> : = <日本語ターム> | <日本語修飾子> / <日本語ターム>.

【0056】<英語ターム> · <日本語ターム> / ... / <日本語ターム>は、英語タームが表す概念のうち、日本語タームが表す概念と共通の概念を指示する。たとえば、「bank・銀行」は、お金に関わる業務を行う組織としてのbankを指示し、「bank・岸」は、川や湖に沿った場所としてのbankを指示する。<英語概念ラベル> + <英語概念ラベル>は、二つの概念ラベルが指示する概念の共通部分を核とし、それぞれの概念ラベルが指示する

$$\begin{aligned} \text{While } \exists y \cdot x_1/x_2/\dots/x_k (\in C_E(y)) \\ \text{and } y \cdot x'_1/x'_2/\dots/x'_k (\in C_E(y)) \end{aligned}$$

概念を合わせた範囲の概念を指示する。「duty・税+tax」、「plane・飛行機+airplane」、及び「reasoning+inference」などが例である。

【0057】英語概念ラベルデータ94は、各英語ターム $y$ に対応する英語概念ラベル集合を表すレコード $C_E(y)$ の集まりである。英語概念ラベル集合の例を示す。

$C_E(\text{interest}) = \{\text{interest} \cdot \text{興味}, \text{interest} \cdot \text{金利/利率}\}.$

このレコード $C_E(\text{interest})$ は、英日ターム結合データ92中に

$TL_{EJ}(\text{interest}) = \{\text{興味}, \text{金利}, \text{利率}\}.$

なるレコードが含まれるとき、それに対応して生成される。

【0058】英語概念の関連タームデータ96は、各英語概念ラベルに対する関連ターム集合を表すレコードの集まりで、以下のように記す。

$$RT_E(Y_i) = \{y(N+i, 1), y(N+i, 2), \dots, y(N+i, n_{N+i})\} \\ (i = 1, 2, \dots, Q).$$

ここで、 $Y$ は英語概念ラベル、 $y$ は英語タームである。

【0059】英語概念の関連タームデータ96は、最終目的である英語概念関連シソーラス62を生成するための中間データである。英語概念関連シソーラス62として必要なのは、関連ターム集合ではなく、関連概念集合である。しかし、全てのタームに対する概念ラベル集合を生成してからでないと、関連概念集合を作成することはできない。そこで、暫定的に関連ターム集合を作成しておき、後続の英語概念関連シソーラス生成306において関連概念集合に変換する。

【0060】英語ターム $y$ に対する英語概念ラベル集合 $C_E(y)$ と、 $C_E(y)$ 中の概念ラベルに対する関連ターム集合を生成するアルゴリズムは次のとおりである。

1) 英語ターム $y$ が少なくとも一つの日本語タームと結合されているとき

i) 英語概念ラベル集合の初期データを作成する。英語ターム $y$ に結合された日本語タームの各々を日本語修飾子とする英語概念ラベルを生成し、その要素とする。すなわち、

$$C_E(y) \leftarrow \{y \cdot x \mid x \in TL_{EJ}(y)\}.$$

ii) 二つの英語概念の類似度が予め定めた閾値 $\alpha$ 以上であるなら、それらを一つの英語概念に統合する処理を可能な限り繰り返す。すなわち、



$$\begin{aligned} \text{s. t. } S(y \cdot x_1/x_2/\cdots/x_k, y \cdot x'_1/x'_2/\cdots/x'_k) \geq \alpha, \\ C_E(y) \leftarrow C_E(y) - \{y \cdot x_1/x_2/\cdots/x_k, y \cdot x'_1/x'_2/\cdots/x'_k\} \\ + \{y \cdot x_1/x_2/\cdots/x_k/x'_1/x'_2/\cdots/x'_k\}. \end{aligned}$$

ここで、共通の英語ターム  $y$  に関わる二つの英語概念  $Y$  【0061】

$$Y_1 = y \cdot x_1/x_2/\cdots/x_k \text{ と } Y_2 = y \cdot x'_1/x'_2/\cdots/x'_k \quad \text{【数2】}$$

の類似度  $S(Y_1, Y_2)$  は次式で定義される。

$$\begin{aligned} S(Y_1, Y_2) = & |(\bigcup_{1 \leq i \leq k} RT_J(x_i)) \cap (\bigcup_{1 \leq i \leq k'} RT_J(x'_i))| \\ & / |(\bigcup_{1 \leq i \leq k} RT_J(x_i)) \cup (\bigcup_{1 \leq i \leq k'} RT_J(x'_i))|. \end{aligned}$$

【0062】すなわち、概念ラベルを構成する日本語修飾子の関連ターム集合間の重なり度で定義される。

iii) 処理ii) の結果として得られた英語概念の各々に対して関連ターム集合データを作成する。英語概念  $y \cdot$  【0063】

$$RT_E(y \cdot x_1/x_2/\cdots/x_k) = RT_E(y)$$

$$- \{ \bigcup_{x'' \in \bigcup_{1 \leq i \leq k} RT_J(x_i)} TL_{JE}(x'') - \bigcup_{x'' \in \bigcup_{1 \leq i \leq k'} RT_J(x'_i)} TL_{JE}(x'') \}.$$

【0064】ここで、 $JM1$  は日本語修飾子の要素の集合である。すなわち、 $JM1 = \{x_1, x_2, \cdots, x_k\}$ 。  $JM2$  は英語ターム  $y$  に結合された日本語タームの集合から日本語修飾子の要素を除いたものである。すなわち、 $JM2 = TL_{EJ}(y) - JM1$ 。

2) 英語ターム  $y$  が日本語ターム  $x$  と結合されていないとき

i) 英語ターム  $y$  そのものを概念ラベルとする。英語概念ラベル集合はこれを唯一の要素とする。すなわち、 $C_E(y) \leftarrow \{y\}$ 。

ii) 英語ターム  $y$  の関連ターム集合  $RT_E(y)$  をそのまま英語概念ラベル  $y$  の関連ターム集合とする。上記アルゴリズムの1)のii)における閾値  $\alpha$  の設定について補足しておく。ここでの目的は、一つの英語タームが表す複数の概念を区別するための日本語修飾子を得ることである。したがって、閾値  $\alpha$  は小さめに設定し、類義の日本語訳語を一つの日本語修飾子に統合するのがよい。

【0065】日本語概念ラベル生成302は、英語概念ラベル生成303と同様である。その出力である日本語概念ラベルデータ93は、英語概念ラベルデータ94と同様で、日本語の各ターム  $x$  に対応する日本語概念ラベル集合を表すレコード  $C_J(x)$  の集まりである。日本語概念ラベルは英語概念ラベルと同様で、以下のようにタームを組み合わせたものである。

【0066】 $\langle \text{日本語概念ラベル} \rangle := \langle \text{日本語ターム} \rangle | \langle \text{日本語ターム} \rangle \cdot \langle \text{英語修飾子} \rangle | \langle \text{日本語概念ラベル} \rangle + \langle \text{日本語概念ラベル} \rangle$ 。

$\langle \text{英語修飾子} \rangle := \langle \text{英語ターム} \rangle | \langle \text{英語修飾子} \rangle / \langle \text{英語ターム} \rangle$ 。

日本語概念ラベル生成302のもう一つの出力である日

$x_1/x_2/\cdots/x_k$  の関連ターム集合  $RT_E(y \cdot x_1/x_2/\cdots/x_k)$  は次式のとおりでである。

【0063】

【数3】

本語概念の関連タームデータ95は、英語概念の関連タームデータ96と同様で、日本語の各概念ラベルに対する関連ターム集合を表すレコードの集まりである。すなわち、

$$RT_J(X_i) = \{x(M+i, 1), x(M+i, 2), \cdots, x(M+i, n_{M+i})\} \quad (i = 1, 2, \cdots, P).$$

ここで、 $X$  は日本語概念ラベル、 $x$  は日本語タームである。

【0067】(4) 日英概念結合304

日英概念結合304は、日本語概念ラベルデータ93と英語概念ラベルデータ94とを入力として、日英概念結合データ63と英日概念結合データ64とを生成する。日本語概念ラベルデータ93と英語概念ラベルデータ94とは説明済みであるので、まず日英概念結合データ63と英日概念結合データ64とについて説明する。

【0068】日英概念結合データ63は、日本語の各概念について、それに結合された英語の概念の集合を表すレコードの集まりである。すなわち、

$$CL_{JE}(X_i) = \{Y'(i, 1), Y'(i, 2), \cdots, Y'(i, q'_i)\} \quad (i = 1, 2, \cdots, P).$$

ここで、 $X$  は日本語概念ラベル、 $Y'$  は英語概念ラベルである。

【0069】同様に、英日概念結合データ64は、英語の各概念について、それに結合された日本語の概念の集合を表すレコードの集まりである。すなわち、

$$CL_{EJ}(Y_i) = \{X'(i, 1), X'(i, 2), \cdots, X'(i, p'_i)\} \quad (i = 1, 2, \cdots, Q).$$

ここで、 $Y$  は英語概念ラベル、 $X'$  は日本語概念ラベルである。

【0070】日英概念結合データ63と英日概念結合データ64とを生成するアルゴリズムは次のとおりであ

る。

1) 全ての日本語概念ラベル  $X = x \cdot y_1 / y_2 / \dots / y_k$  に

$$CL_{JE}(X) = \{Y \mid Y = y \cdot x_1 / x_2 / \dots / x_k (\in C_E(y)), y \in \{y_1, y_2, \dots, y_k\}, \{x_1, x_2, \dots, x_k\} \ni x\}.$$

すなわち、 $X$  の英語修飾子に含まれる英語ターム  $y$  の英語概念集合  $C_E(y)$  の要素  $Y$  であって、日本語修飾子に  $x$  を含むものの集合を生成する。

$$CL_{EJ}(Y) = \{X \mid X = x \cdot y_1 / y_2 / \dots / y_k (\in C_J(x)), x \in \{x_1, x_2, \dots, x_k\}, \{y_1, y_2, \dots, y_k\} \ni y\}.$$

すなわち、 $Y$  の日本語修飾子に含まれる日本語ターム  $x$  の日本語概念集合  $C_J(x)$  の要素  $X$  であって、英語修飾子に  $y$  を含むものの集合を生成する。

【0072】日英概念結合304の出力例を示す。日本語概念ラベルデータ93が

$C_J(\text{興味}) = \{\text{興味} \cdot \text{interest}\}$ 、

$C_J(\text{金利}) = \{\text{金利} \cdot \text{interest}\}$ 、

$C_J(\text{利率}) = \{\text{利率} \cdot \text{interest}\}$

であり、英語概念ラベルデータ94が

$C_E(\text{interest}) = \{\text{interest} \cdot \text{興味}, \text{interest} \cdot \text{金利/利率}\}$

であるとする。このとき、日英概念結合データ63として

$CL_{JE}(\text{興味} \cdot \text{interest}) = \{\text{interest} \cdot \text{興味}\}$ 、

$CL_{JE}(\text{金利} \cdot \text{interest}) = \{\text{interest} \cdot \text{金利/利率}\}$ 、

$CL_{JE}(\text{利率} \cdot \text{interest}) = \{\text{interest} \cdot \text{金利/利率}\}$

が生成され、英日概念結合データ64として

$CL_{EJ}(\text{interest} \cdot \text{興味}) = \{\text{興味} \cdot \text{interest}\}$ 、

$CL_{EJ}(\text{interest} \cdot \text{金利/利率}) = \{\text{金利} \cdot \text{interest}, \text{利率} \cdot \text{interest}\}$

が生成される。

【0073】ここで、概念ラベルの表記に関する一つの規則を定める。ある英語タームに対応する英語概念がただ一つであるならば、日本語修飾子をつけることは無意味であり、タームそのものを概念ラベルとして差し支えない。日本語タームに関しても同様である。上に述べた日英概念結合304のアルゴリズムと違って、これ以降の処理では、日本語修飾子や英語修飾子に基づく判断を含まない。したがって、タームの唯一の概念である概念については、概念結合データ63、64の出力時に、概念ラベルをタームそのものに変更することにする。この規則に従えば、上の例における日英概念結合データ63は

$CL_{JE}(\text{興味}) = \{\text{interest} \cdot \text{興味}\}$ 、

$CL_{JE}(\text{金利}) = \{\text{interest} \cdot \text{金利/利率}\}$ 、

$CL_{JE}(\text{利率}) = \{\text{interest} \cdot \text{金利/利率}\}$

となり、英日概念結合データ64は

$CL_{EJ}(\text{interest} \cdot \text{興味}) = \{\text{興味}\}$ 、

$CL_{EJ}(\text{interest} \cdot \text{金利/利率}) = \{\text{金利}, \text{利率}\}$

となる。ここでは、「興味」が単一の概念を表す語であ

対して、

【0071】2) 全ての英語概念ラベル  $Y = y \cdot x_1 / x_2 / \dots / x_k$  に対して、

るので、概念ラベル「興味・interest」が「興味」に略記されている。「金利・interest」「利率・interest」も同様で、それぞれ「金利」「利率」に略記されている。一方、「interest」は複数の概念を表す語であるので、概念ラベル「interest・興味」「interest・金利/利率」を略記することはできない。

【0074】(5) 日本語概念関連シソーラス生成305

(6) 英語概念関連シソーラス生成306

日本語概念関連シソーラス生成305は、日本語概念ラベルデータ93と日本語概念の関連タームデータ95とを入力して、日本語概念関連シソーラス61を出力する。英語概念関連シソーラス生成306は、英語概念ラベルデータ94と英語概念の関連タームデータ96とを入力として、英語概念関連シソーラス62を出力する。これらの入力については説明済みである。

【0075】出力である日本語概念関連シソーラス61と英語概念関連シソーラス62は次のとおりである。日本語概念関連シソーラス61は、日本語の各概念の関連概念集合を表すレコードの集まりである。すなわち、 $RC_J(X_i) = \{X(i, 1), X(i, 2), \dots, X(i, p_i)\}$  ( $i = 1, 2, \dots, P$ )。

ここで、 $X$  は日本語概念ラベルである。関連概念集合の例を以下に示す。

【0076】 $RC_J(\text{銀行}) = \{\text{ローン}, \text{金利}, \text{口座}, \text{利率}, \text{証券}, \text{経済}, \text{金融}, \text{投資}\}$ 。

$RC_J(\text{岸}) = \{\text{川}, \text{水}, \text{ボート}, \text{湖}, \text{釣り}\}$ 。

英語概念関連シソーラス62は、英語の各概念の関連概念集合を表すレコードの集まりである。すなわち、

$RC_E(Y_i) = \{Y(i, 1), Y(i, 2), \dots, Y(i, q_i)\}$  ( $i = 1, 2, \dots, Q$ )。

ここで、 $Y$  は英語概念ラベルである。関連概念集合の例を以下に示す。

$RC_E(\text{bank} \cdot \text{銀行}) = \{\text{account} \cdot \text{口座}, \text{interest} \cdot \text{金利/利率}, \text{loan}, \text{investment}, \text{economy}\}$ 。

$RC_E(\text{bank} \cdot \text{岸}) = \{\text{river}, \text{boat}, \text{water}, \text{fishing}, \text{park} \cdot \text{公園}, \text{lake}\}$ 。

【0077】英語概念関連シソーラス生成306のアルゴリズムは以下のとおりである。なお、日本語概念関連

シソーラス生成305のアルゴリズムも全く同様である。英語概念Yの関連ターム集合 $RT_E(Y)$ の各要素yに対応して、yの概念ラベル集合 $C_E(y)$ の要素のうちYとの相関度が最大のものを関連概念集合 $RC_E(Y)$ の

$$RC_E(Y) = \{Y' \mid Y \in RT_E(Y), Y' \in C_E(y), S_2(Y', Y) = \max_{Y'' \in C_E(y)} S_2(Y'', Y)\}.$$

【0079】ここで、 $S_2$ は関連ターム集合に基づく英語概念の相関度で、次式で定義される。

$$S_2(Y_1, Y_2) = |RT_E(Y_1) \cap RT_E(Y_2)| / |RT_E(Y_1) \cup RT_E(Y_2)|.$$

たとえば、英語概念「bank・銀行」の関連ターム集合が「interest」を含み、英語ターム「interest」の概念ラベル集合が{interest・興味, interest・金利/利率}であるとする。このとき、「bank・銀行」と「interest・興味」との相関度、「bank・銀行」と「interest・金利/利率」との相関度が計算される。後者の相関度が大きければ、「bank・銀行」の関連概念集合の要素として「interest・金利/利率」が選択される。

【0080】(7) 日本語概念マージ307

(8) 英語概念マージ308

日本語概念マージ307と英語概念マージ308の処理は、日本語と英語の役割が反転する以外、全く同様である。したがって、ここでは英語概念マージ308について説明する。

【0081】英語概念マージ308は、日本語の同一概念に結合された英語概念で類似度の高いものをマージして一つ概念にする。入力、英語概念関連シソーラス62、日英概念結合データ63、英日概念結合データ64であり、出力はそれらの更新データである。

【0082】英語概念マージ308のアルゴリズムは以下のとおりである。全ての日本語概念Xに関して、以下の処理を可能な限り繰り返す。 $Y_1, Y_2 \in CL_{JE}(X)$ で $S_3(Y_1, Y_2) \geq \beta$ なる英語概念の組 $Y_1, Y_2$ が存在するならば、a) からc) を実行する。ここで、 $S_3(Y_1, Y_2)$ は英語概念 $Y_1$ と $Y_2$ の類似度で、次式で定義される。 $S_3(Y_1, Y_2) = |RC_E(Y_1) \cap RC_E(Y_2)| / |RC_E(Y_1) \cup RC_E(Y_2)|.$

【0083】a) 英語概念関連シソーラス62の更新  
全ての $Y \in RC_E(Y_1)$ に関して、 $RC_E(Y) \leftarrow RC_E(Y) - \{Y_1\} + \{Y_1 + Y_2\}.$

全ての $Y \in RC_E(Y_2)$ に関して、 $RC_E(Y) \leftarrow RC_E(Y) - \{Y_2\} + \{Y_1 + Y_2\}.$

$RC_E(Y_1 + Y_2) \leftarrow RC_E(Y_1) \cup RC_E(Y_2).$

$RC_E(Y_1)$ と $RC_E(Y_2)$ を消去する。

【0084】b) 日英概念結合データ63の更新  
全ての $x \in CL_{EJ}(Y_1)$ に関して、 $CL_{JE}(X) \leftarrow CL_{JE}(X) - \{Y_1\} + \{Y_1 + Y_2\}.$

全ての $X \in CL_{EJ}(Y_2)$ に関して、 $CL_{JE}(X) \leftarrow CL_{JE}(X) - \{Y_2\} + \{Y_1 + Y_2\}.$

要素として選択する。すなわち、

【0078】

【数4】

$$CL_{JE}(X) \leftarrow \{Y_2\} + \{Y_1 + Y_2\}.$$

c) 英日概念結合データ64の更新

$$CL_{EJ}(Y_1 + Y_2) \leftarrow CL_{EJ}(Y_1) \cup CL_{EJ}(Y_2).$$

$CL_{EJ}(Y_1)$ と $CL_{EJ}(Y_2)$ を消去する。

【0085】上記アルゴリズム中の閾値 $\beta$ は大きめに設定し、類似度が非常に高い概念のみをマージするのがよい。概念の範囲やニュアンスが異なるタームを別々のエンティティとするほうが、利用価値の高いシソーラスになるからである。この点は、相手言語のタームが表す複数の概念を区別することが目的の場合（英語概念ラベル生成303のアルゴリズムにおける閾値 $\alpha$ ）と異なっている。

【0086】以上説明した処理によって、日本語コーパス51と英語コーパス52が対になった日英二言語テキストコーパスから、日本語概念関連シソーラス61、英語概念関連シソーラス62、日英概念結合データ63、英日概念結合データ64からなる日英パラレルシソーラスを生成することができる。このように生成された日英パラレルシソーラスは、図1に示す通信ネットワーク3を介して、クライアント計算機2によるシソーラスナビゲーションに利用される。つぎに、図5～図7を用いて本システムの処理を説明する。

【0087】図5は、本システムにおいて、クライアント計算機2の表示画面の内容を説明する図である。図5に示す表示画面は、概念集合エリア1010、ズームインエリア1020及び機能選択ボタンから構成される。機能選択ボタンには、ズームインボタン1030、翻訳ボタン1040、クリアボタン1050、終了ボタン1060がある。

【0088】ズームインエリア1020には、一つ以上の概念クラスタ1021がそれに対応付けられた選択ボタン1022とともに表示される。この概念クラスタ1021は、関連性の高い概念の集合である。たとえば、「地球環境問題」に該当する概念クラスタであれば、「地球温暖化」、「オゾン層」、「温室効果」、「フロン」及び「大気」等の概念が表示される。クライアント計算機2のユーザは、これら概念クラスタ1021を複数指定することができる。

【0089】図6は、本実施の形態によるパラレルシソーラスナビゲーションシステムの処理を説明するフローチャートである。以下、図5に示した表示内容と対応して本システムの処理を説明する。最初に初期画面を表示

する(ステップ410)。初期画面では、概念集合エリア1010とズームインエリア1020は空白である。本システムには、「日本語」/「英語」を切り替える言語インジケータが内部に設けられており、初期画面を表示したときには、言語インジケータを「日本語」にする。

【0090】初期画面表示(ステップ410)の後、入力待ちの状態になる(ステップ420)。この状態では、概念集合エリア1010は書き込み可能であり、通常、ユーザが一つ以上の日本語タームあるいは日本語概念ラベルを書き込む。入力待ちの状態では押されたボタンにより、以下のように分岐する。

【0091】(1)ズームインボタン1030が押されたとき

概念集合エリア1010に表示されている概念集合を読み込む(ステップ430)。ユーザが概念集合エリア1010に書き込むのは、通常、概念ラベルでなくタームである。タームが書き込まれている場合には、該タームから生成された全ての概念ラベルが書き込まれているとみなして処理する。この処理は、言語インジケータが「日本語」のときには日本語概念関連シソーラス61を参照し、言語インジケータが「英語」のときには英語概念関連シソーラス62を参照することにより行われる。

【0092】つぎに、概念集合に含まれる概念と関連の強い概念を加えて概念集合を拡大し、拡大された概念集合をクラスタリングする(ステップ440)。この処理は、言語インジケータが「日本語」のときには日本語概念関連シソーラス61を参照し、言語インジケータが「英語」のときには英語概念関連シソーラス62を参照することにより行われる。最後に、得られた概念クラスを選択ボタン1022とともにズームインエリア1020に表示し(ステップ450)、入力待ちの状態に戻る。

【0093】(2)概念クラスタの選択ボタン1022が押されたとき

選択された概念クラスタ1021を概念集合エリア1010にコピー(上書き)し(ステップ460)、入力待ちの状態に戻る。入力待ちの状態では、概念集合エリア1010は書き込み可能であり、ユーザがタームあるいは概念ラベルを追加したり、削除したりすることが可能である。

【0094】(3)翻訳ボタン1040が押されたとき概念集合エリア1010に表示されている概念集合を読み込む(ステップ470)。この処理はステップ430と全く同じである。つぎに、概念集合を翻訳する(ステップ480)。言語インジケータが「日本語」のときには日本語概念集合から英語概念集合への翻訳が実行され、言語インジケータが「英語」のときには英語概念集合から日本語概念集合への翻訳が実行される。

【0095】最後に、翻訳結果を概念集合エリア101

0に表示(上書き)し、言語インジケータを反転させ(ステップ490)、入力待ちの状態に戻る。入力待ちの状態では、概念集合エリア1010は書き込み可能であり、ユーザがタームあるいは概念ラベルを追加したり、削除したりすることが可能である。

【0096】(4)クリアボタン1050が押されたとき

初期画面表示状態(ステップ410)に戻る。

(5)終了ボタン1060が押されたとき  
処理を終了する。

以上述べた処理により、言語間の遷移を含むパラレルシソーラスのナビゲーションが可能になる。

【0097】図7は、本システムを特徴付ける概念集合翻訳(ステップ480)の処理を詳細に説明する図である。図7は、日本語概念集合を英語概念集合に翻訳する処理を示したものであるが、英語概念集合を日本語概念集合に翻訳する処理も全く同様である。

【0098】入力日本語概念集合が与えられると、日英概念結合データ63を参照して、日本語概念集合中の日本語概念に結合されている英語概念を集めて、核となる英語概念集合を生成する(ステップ481)。

【0099】つぎに、英語概念関連シソーラス62を参照して、この英語概念集合に含まれる英語概念と関連の強い英語概念を集め、さらに、日英概念結合データ64を参照して、関連英語概念のうちで日本語概念と結合されていないものを選択する。核となる英語概念集合に選択した英語概念を追加して翻訳結果とする(ステップ482)。

【0100】日本語概念集合から英語概念集合への翻訳(遷移)の例を示す。入力の日本語概念集合は{地球温暖化, オゾン層, 温室効果, フロン, 大気, 二酸化炭素, 環境}であるとする。日英概念結合データ63は以下のレコードを含むとする。

【0101】 $CL_{JE}(\text{地球温暖化}) = \phi$ .

$CL_{JE}(\text{オゾン層}) = \{\text{ozone layer}\}$ .

$CL_{JE}(\text{温室効果}) = \phi$ .

$CL_{JE}(\text{フロン}) = \phi$ .

$CL_{JE}(\text{大気}) = \{\text{atmosphere} \cdot \text{大気}\}$ .

$CL_{JE}(\text{二酸化炭素}) = \{\text{carbon dioxide}\}$ .

$CL_{JE}(\text{環境}) = \{\text{environment}\}$ .

さらに、英語概念関連シソーラス62が以下のレコードを含むとする。

【0102】 $RC_J(\text{ozone layer}) = \{\text{chlorofluorocarbon, depletion, atmosphere} \cdot \text{大気, warming}\}$ .

$RC_J(\text{atmosphere} \cdot \text{大気}) = \{\text{pollution, environment, gas} \cdot \text{気体/ガス, carbon dioxide}\}$ .

$RC_J(\text{carbon dioxide}) = \{\text{atmosphere} \cdot \text{大気, energy, warming, environment, regulation}\}$ .

$RC_J(\text{environment}) = \{\text{protection, carbon dioxide, energy, atmosphere} \cdot \text{大気, pollution}\}$ .

また、英日概念結合データ64は以下のレコードを含むとする。

【0103】CL<sub>EJ</sub>(ozone layer)={オゾン層}.

CL<sub>EJ</sub>(chlorofluorocarbon)=φ.

CL<sub>EJ</sub>(depletion)={破壊}.

CL<sub>EJ</sub>(atmosphere・大気)={大気}.

CL<sub>EJ</sub>(warming)=φ.

CL<sub>EJ</sub>(pollution)={汚染}.

CL<sub>EJ</sub>(environment)={環境}.

CL<sub>EJ</sub>(gas・気体/ガス)={気体, ガス}.

CL<sub>EJ</sub>(carbon dioxide)={二酸化炭素}.

CL<sub>EJ</sub>(energy)={エネルギー}.

CL<sub>EJ</sub>(regulation)={規制}.

CL<sub>EJ</sub>(protection)={保護}.

【0104】このとき、日本語概念集合{地球温暖化, オゾン層, 温室効果, フロン, 大気, 二酸化炭素, 環境}から英語概念集合への翻訳結果は{ozone layer, atmosphere・大気, carbon dioxide, environment, chlorofluorocarbon, warming}になる。

【0105】英語概念集合を構成する6つの英語概念のうち、「ozone layer」、「atmosphere・大気」、「carbon dioxide」及び「environment」の4つは、日本語概念集合中の日本語概念とシソーラス中で陽に結合されていたものである。また、「chlorofluorocarbon」及び「warming」の2つは、日本語概念集合中の日本語概念とシソーラス中で陽に結合されていなかったが、上記4つの英語概念の関連概念として追加されたものである。実は、「chlorofluorocarbon」は「フロン」の英語訳であり、「warming」は「地球温暖化」の英語訳の一部である。このようにして、概念結合として陽に表現されていない対訳を含む翻訳結果を得ることができる。

【0106】以上、この発明の実施の形態を図面を参照して詳述してきたが、具体的な構成はこれらの実施の形態に限られるものではなく、この発明の要旨を逸脱しない範囲の設計の変更等があってもよい。はじめに、上記実施の形態では、日本語と英語とのパラレルシソーラスを生成しているが、たとえば、日本語とフランス語、更には、日本語のものも含めて一般的な2ヶ国語によるパラレルシソーラスを生成するものであってもよい。

【0107】また、上記実施の形態では、図4に示したように、多義語と同義語の各々が有する概念を考慮した機能を実現しているが、多義語の概念のみを考慮した機能、又は同義語のみを考慮した機能を実現してもよい。この場合、日本語概念マージ307及び英語概念マージ308の機能を選択的に設けることで実現できる。

【0108】また、本発明におけるクライアント計算機2としては、有線回線により通信ネットワーク3に接続されるパーソナルコンピュータ又はワークステーション等、また、無線回線により通信ネットワーク3に接続される移動体通信端末(携帯電話、PHS(Personal Han-

dy-Phone System)、PDA(Personal Digital Assistance)等)であつてもよい。

【0109】なお、本発明のパラレルシソーラスの生成装置、及びパラレルシソーラスナビゲーションシステムは、このサーバ計算機1又はクライアント計算機2を機能させるためのプログラムによっても実現される。このプログラムは、たとえばCD-ROM等のコンピュータで読み取り可能な記録媒体に格納されている。

【0110】パラレルシソーラスの生成装置、又はパラレルシソーラスナビゲーションシステムを機能させるためのプログラムを記録した記録媒体は、図1に示す記憶装置13そのものであつてもよいし、また、外部記憶装置としてCD-ROMドライブ等のプログラム読み取り装置(図示せず)が設けられ、そこに挿入することで読み取り可能なCD-ROM等であつてもよい。また、上記記録媒体は、磁気テープ、カセットテープ、フロッピーディスク、ハードディスク、MO/MD/DVD等、又は半導体メモリであつてもよい。

【0111】また、本発明のパラレルシソーラスの生成装置により生成されたパラレルシソーラスは、CD-ROM等のコンピュータで読み取り可能な記録媒体に格納されてもよい。このパラレルシソーラスは、二つの言語のシソーラスを結合したものであり、日本語及び英語概念ラベル生成302、303により二つの言語のタームを組み合わせた概念ラベルが生成され、日英概念結合304(図4参照)により、二つの言語の概念ラベルが概念に基づいて結合されている。

【0112】

【発明の効果】本発明のパラレルシソーラスの生成プログラムを記録した記録媒体によれば、二つの言語の関連シソーラスが結合されたパラレルシソーラスを二つの言語のテキストコーパスから自動的に生成することができる。生成されるパラレルシソーラスは、概念と概念の関連を示したものである。

【0113】また、本発明のパラレルシソーラスを記録した記録媒体によれば、タームとタームの関連を示す従来のシソーラスと異なり、多義語や同義語の問題が解決されているので、本発明により生成されたシソーラスを用いることで、各種の自然言語処理システムの精度を向上することができる。

【0114】また、本発明のパラレルシソーラスナビゲーションプログラムを記録した記録媒体によれば、複数言語にまたがる効率的なテキストマイニングが可能になる。特に、母国語のシソーラスをナビゲーションして外国語の情報にアクセスすることが容易になる。従来のクロスランゲージ情報検索において問題とされる検索要求の翻訳精度も、概念集合の遷移(翻訳)機能により大きく改善される。

【図面の簡単な説明】

【図1】本発明の実施の形態によるパラレルシソーラス

の生成装置と、該装置を収容するパラレルシソーラスナビゲーションシステムの構成を説明するブロック図である。

【図2】本発明の実施の形態によるパラレルシソーラスの生成装置における入出力とモジュール構成を機能的に説明する図である。

【図3】日本語シソーラス生成の処理の詳細を説明する図である。

【図4】日英シソーラス結合モジュールの処理の詳細を説明する図である。

【図5】本発明の実施の形態によるパラレルシソーラスナビゲーションシステムにおいて、クライアント計算機の表示画面の内容を説明する図である。

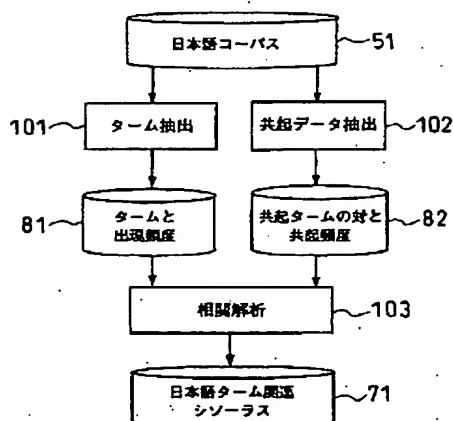
【図6】本発明の実施の形態によるパラレルシソーラスナビゲーションシステムの処理を説明するフローチャートである。

【図7】本発明の実施の形態によるパラレルシソーラスナビゲーションシステムを特徴付ける概念集合翻訳（ステップ480）の処理を詳細に説明する図である。

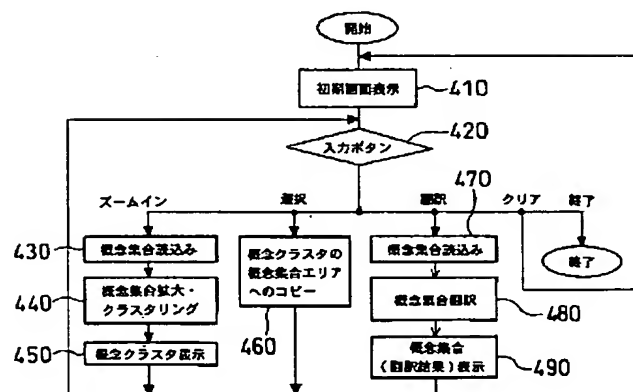
#### 【符号の説明】

- |    |              |      |                |
|----|--------------|------|----------------|
| 1  | サーバ計算機       | 63   | 日英概念結合データ      |
| 2  | クライアント計算機    | 64   | 英日概念結合データ      |
| 3  | 通信ネットワーク     | 71   | 日本語ターム関連シソーラス  |
| 10 | 日本語シソーラス生成   | 72   | 英語ターム関連シソーラス   |
| 11 | 処理装置         | 73   | 日英対訳辞書         |
| 12 | 入力装置         | 74   | 英日対訳辞書         |
| 13 | 記憶装置         | 81   | タームと出現頻度       |
| 20 | 英語シソーラス生成    | 82   | 共起タームの対と共起頻度   |
| 30 | 日英シソーラス結合    | 91   | 日英ターム結合データ     |
| 51 | 日本語コーパス      | 92   | 英日ターム結合データ     |
| 52 | 英語コーパス       | 93   | 日本語概念ラベルデータ    |
| 61 | 日本語概念関連シソーラス | 94   | 英語概念ラベルデータ     |
| 62 | 英語概念関連シソーラス  | 95   | 日本語概念の関連タームデータ |
|    |              | 96   | 英語概念の関連タームデータ  |
|    |              | 101  | ターム抽出          |
|    |              | 102  | 共起データ抽出        |
|    |              | 103  | 相関解析           |
|    |              | 301  | 日英ターム結合        |
|    |              | 302  | 日本語概念ラベル生成     |
|    |              | 303  | 英語概念ラベル生成      |
|    |              | 304  | 日英概念結合         |
|    |              | 305  | 日本語概念関連シソーラス生成 |
|    |              | 306  | 英語概念関連シソーラス生成  |
|    |              | 307  | 日本語概念マージ       |
|    |              | 308  | 英語概念マージ        |
|    |              | 1010 | 概念集合エリア        |
|    |              | 1020 | ズームインエリア       |
|    |              | 1021 | 概念クラス          |
|    |              | 1022 | 選択ボタン          |
|    |              | 1030 | ズームインボタン       |
|    |              | 1040 | 翻訳ボタン          |
|    |              | 1050 | クリアボタン         |
|    |              | 1060 | 終了ボタン          |

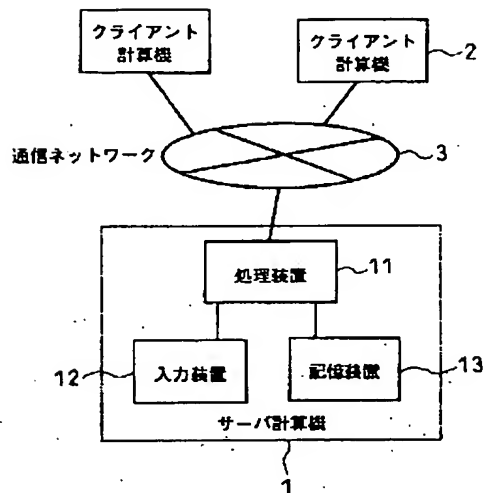
【図3】



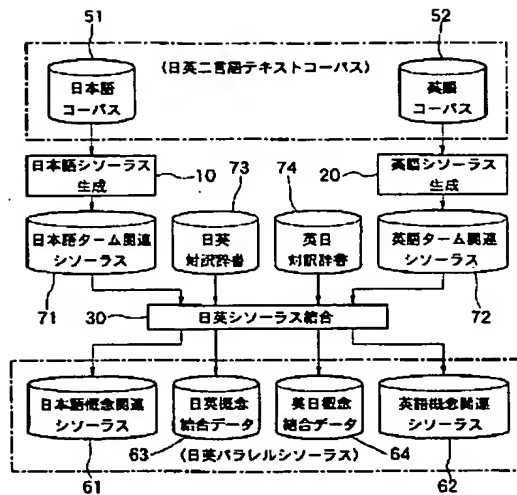
【図6】



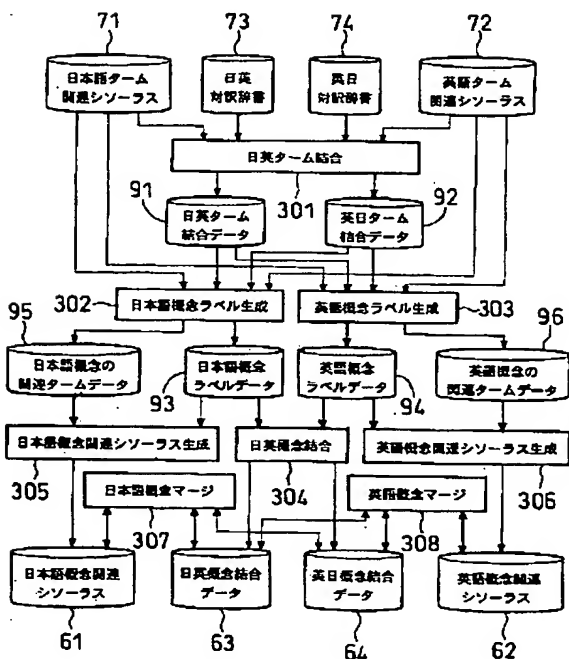
【図 1】



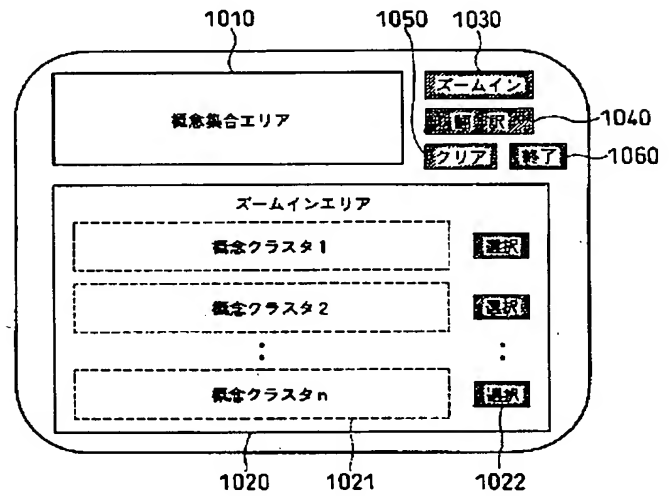
【図 2】



【図 4】



【図 5】



【図 7】

